A SURVEY ON INFREQUENT PATTERN MINING

Kamepalli Sujatha and K. Raja Sekhara Rao,
[1]Research Scholar, Department of CSE, Krishna University,
Machilipatnam, Krishna District, Andhra Pradesh, India
[2]Dean Administration, Professor, Department of CSE,
K. L. University, Guntur District, Andhra Pradesh, India

ABSTRACT
An infrequent pattern is an item set or a rule whose support is less than the minimum support threshold. The extraction of infrequent patterns is called infrequent pattern mining. This work mainly concentrates on the infrequent pattern mining. It gives a survey on methods for mining infrequent patterns from different types of data set. This paper reviews different research papers and presents the methods that they adopted to mine the infrequent patterns. It also explains about different application areas where these infrequent patterns can be used.

KEYWORDS: *Minimum Support, Infrequent Patterns, Frequent Patterns, pattern mining.*

I. INTRODUCTION

The mining task that focuses on discovering frequent patterns from the databases is called frequent pattern mining. In frequent pattern mining, only frequent patterns are returned while infrequent patterns are simply discarded without further consideration. This is because the most valuable information is carried by the frequent patterns and the infrequent patterns cannot adequately reflect the typical characteristics from the data because of their rare occurrence [5]. However, since the late 1990s, more and more researchers have realized the importance of infrequent patterns with the increasing demands from applications of anomaly detection, especially in medicine [3], genetics [7], molecular biology [4] and network security [8]. In these areas, infrequent patterns are considered significant due to the huge influence they may have. In the study of finding a better treatment approach for a special disease, researchers would like spend more time on studying an abnormal case rather than reading the millions of records of healthy people [9].

In this scenario, more effort has been put into the development of infrequent pattern mining [6]. Frequent patterns are item sets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent item set. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as sub graphs, sub trees, or sub lattices, which may be combined with item sets or subsequences. If a substructure occurs frequently in a graph database, it is called a (frequent) structural pattern. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data [12].

Often when considering data mining, the focus is on frequent patterns. Although the majority of the most interesting patterns will lie within the frequent ones, there are important patterns that will be ignored with this approach. These are called infrequent patterns. Take for example the sale of VHS:s and DVD:s here will be low occurrences of people buying both of them. In terms of data mining, the item set {VHS, VD} will be infrequent and therefore ignored. However, people that buys DVD:s does not tend to buy VHS:s and vice versa. These items will be competing and an interesting pattern is found.

Mining infrequent patterns is a challenging endeavor because there is an enormous number of such patterns that can be derived from a given data set. More specifically, the key issues in mining infrequent patterns are: (1) how to identify interesting infrequent patterns, and (2) how to efficiently discover them in large data sets [11]. To get a different perspective on various types of interesting infrequent patterns, two related concepts are negative patterns and negatively correlated patterns.
The remaining part of the paper is organized as it gives the definitions for negative patterns, negative item set, and negative association rule. In the next sections it explains research challenges, review of different papers, applications, results and discussions, conclusion, future work and finally the references.

II. NEGATIVE PATTERNS

Let I = {*i*1, *i*2,…, *id*} be a set of items. A negative item, *ik*, denotes the absence of item *ik* from a given transaction. For examples, coffee is a negative item whose value is 1 if a transaction does not contain coffee.

2.1. Negative Item Set

A negative item set *X* is an item set that has the following properties:
(1) $X = A \cup \acute{B}$, where A is a set of positive items, $\acute{B}$ is a set of negative items, $|\acute{B}| \geq 1$, and
(2) $s(X) \geq minsup$.

2.2. Negative Association Rule

A negative association rule is an association rule that has the following properties: (1) the rule is extracted from a negative item set, (2) the support of the rule is greater than or equal to *minsup*, and (3) the confidence of the rule is greater than or equal to minconf.
The negative item sets and negative association rules are collectively known as negative patterns. An example of negative association rule is tea → coffee', which may suggest that people who drink tea tend to not drink coffee.
As an important research topic in data mining, pattern mining aims to discover unknown or hidden patterns from a large-scale collection of data. Given a database, it is not hard to find some patterns. However, to find the most interesting and useful data patterns highly relevant to the user's application targets is where the challenge comes from. In addition, if the data is represented in complex structures, then it is even more difficult to accomplish the task efficiently. Many of the researchers do work on frequent pattern mining. However, infrequent pattern mining and its applications is still an open topic that has been done a little work previously [1]. I would like to move on surveying the efficient and effective mining algorithms in infrequent pattern mining which explains about infrequent item set mining. Infrequent subsequence mining, infrequent sub structure (sub tree) mining.

III. RESEARCH CHALLENGES

Even though infrequent pattern mining is still an emerging research field and has been studied for a decade, there are still many unsolved topics that can be explored, such as how to further control the number of generated candidate and how to improve the efficiency of the mining process by providing more targeted candidates, etc [2]. It is not hard to generate a large number of infrequent sub trees, but the challenge comes from how to determine which infrequent sub trees are most interesting and valuable to end users and their applications [1].

IV. REVIEW OF DIFFERENT PAPERS

Ling Zhou at all 2007 in "Efficient association rule mining among both frequent and infrequent items" proposed Matrix-based scheme (MBS) and Hash-based scheme (HBS) for mining both frequent and infrequent patterns [13].
Mehdi Adda at all 2012 in "pattern detection with rare item-set mining" explained about rare item sets and non-present item sets and also explained an apriori based method for mining rare item sets and non-present item sets[14].

Alex tze hiang sim at all 2008 in Mining Infrequent and Interesting Rules from Transaction Records" proposed Proportional Error Reduction Technique [15].
Laszlo Szathmary at all 2010 in "Generating Rare Association Rules Using the Minimal Rare Item sets Family" proposed a naïve approach for finding mRIs[16].
Luigi Troiano at all 2009 in" A Fast Algorithm for Mining Rare Item sets" proposed an algorithm for mining rare item sets called Rarity Algorithm[17].
Laszlo Szathmary at all 2012.in "Efficient Vertical Mining of Minimal Rare Item sets" proposed Talky-G and Walky-G algorithms [18].
Ashish Gupta at all 2011 in "Minimally Infrequent Item set Mining using Pattern-Growth Paradigm and Residual Trees" explained the Pattern-Growth Paradigm and Residual Trees for Minimally Infrequent Item set Mining [19].
Budhaditya Saha at all 2007 in "Infrequent Item Mining in Multiple Data Streams" proposed Entropy Based Window Selection method for extracting the Infrequent Patterns from the Data Stream [20].
Xindong wu at all 2004 in "Efficient Mining of Both Positive and Negative Association Rules" explained about the extraction of Positive and Negative Association Rules [21].
Ahmedur Rahman at all in "WiFi Miner: An Online Apriori-Infrequent Based Wireless Intrusion Detection System" proposed an apriori based infrequent pattern mining algorithm for wireless intrusion detection [25].

V. APPLICATIONS

Infrequent patterns can be used in many applications.
In text mining, indirect associations can used to find synonyms, antonym or words that are used in different contexts. For example, the word *data* might be indirectly associated with the word *gold*, using the mediator *mining*.
In the market basket domain, indirect associations can be used to find competing items, such as *desktop computers* and *laptops*, which states that people whom buys desktop computers won't buy laptops.
Infrequent patterns can be used to detect errors. For example, if *{Fire = Yes}* is frequent, but *{Fire = Yes, Alarm = On}* is infrequent, then the alarm system probably is faulting [22].
When evaluating Weka [23], we could not find any signs of an infrequent pattern classifier.
Searching for outliers in data stream is an important area of research in the world of data mining with numerous applications, including credit card fraud detection, discovery of criminal activities in electronic commerce, weather prediction, marketing and customer segmentation [24].The outliers can be detected from infrequent patterns.
Intrusion detection in wireless networks has become a vital part in wireless network security systems with wide spread use of Wireless Local Area Networks (WLAN). Intrusion detection can be done by the infrequent patterns [25].

VI. RESULTS AND DISCUSSIONS

Infrequent pattern mining is still an emerging research field and has been studied for a decade, there are still many unsolved topics that can be explored, such as how to further control the number of generated candidate and how to improve the efficiency of the mining process by providing more targeted candidates, etc. It is not hard to generate a large number of infrequent sub trees, but the challenge comes from how to determine which infrequent sub trees are most interesting and valuable to end users and their applications. By reading this paper one can get a basic knowledge on infrequent patterns and this work acts as a basis for the future work that has been done on this area.

VII. CONCLUSION

This paper mainly concentrated on infrequent patterns. To get a different perspective on various types of interesting infrequent patterns, two related concepts are negative patterns and negatively correlated patterns. It explains what are negative patterns, negative item set, and negative association rule. This paper makes review on different papers related to infrequent patterns and rare item sets and also gives the knowledge on different algorithms proposed for mining infrequent patterns. This also explains

about different application areas where these infrequent patterns are used. This research work elaborates the algorithms for mining infrequent patterns and which becomes the basis for the future work going to be done in this area.

VIII. FUTURE WORK

Mining infrequent patterns is a challenging endeavor because there is an enormous number of such patterns that can be derived from a given data set. More specifically, the key issues in mining infrequent patterns are:
(1) How to identify infrequent patterns,
(2) How to identify the interestingness of those patterns
(3) How to efficiently discover them in large data sets.
To get a different perspective on various types of interesting infrequent patterns, two related concepts are negative patterns and negatively correlated patterns.

REFERENCES

[1] Jia Rong Bit (hons) Advanced Pattern Mining for Complex Data Analysis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy, Deakin University, August 2012.
[2] G. Li, R. Law, J. Rong, and H.Q. Vu. Incorporating both positive and negative association rules into the analysis of outbound tourism in hong kong. Journal of Travel and Tourism Marketing, 27(8):812–828, 2010.
3] F. Medici, M.I. Hawa, A. Giorgini, A. Panelo, C.M. Solfelix, R.D.G. Leslie, and P. Pozzilli. Antibodies to GAD65 and a tyrosine phosphatase-like molecule IA-2ic in Filipino Type I diabetic patients. Diabetes Care, 22(9):1458–1461, 1999.
[4] W. Shi, F.K. Ngok, and D.R. Zusman. Cell density regulates cellular reversal frequency in Myxococcus xanthus. In Proceedings of the National Academy of Sciences of the United States of America, volume 93(9), pages 4142–4146, 1996.
[5] R. Agrawal, T. Imieinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, Proceedings of the ACM SIGMOD International Conference on the Management of Data, pages 207–216, Washington DC, 1993. ACM Press.
[6] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. ACM Transactions on Information Systems, 22(3):381–405, 2004.
[7] A. Sehgal, A. Presente, L. Dudus, and J.F. Engelhardt. Isolation of differentially expressed DNAs during ferret tracheal development: Application of differential display PCR. Experimental Lung Research, 22(4):419–434, 1996.
[8] B. Saha, M. Lazarescu, and S Venkatesh. Infrequent item mining in multiple data streams. In Proceedings of IEEE International Conference on Data Mining (ICDM 2007), pages 569–574, Omaha, NE, October 2007.
[9] J. Yang and J. Logan. A data mining and survey study on diseases associated with paraesophageal hernia. In AMIA Annual Symposium Proceedings, pages 829–833, 2006.
[10] Johanarnle, Peterzhu "Mining Infrequent Patterns" Linköping University, 2009 ,data mining.
[11] Pang-Ning Tan, Michael Steinbach, Vipin Kumar Introduction to data mining, Pearson Education, book.
[12] Jiawei Han · Hong Cheng · Dong Xin · Xifeng Yan. Frequent pattern mining: current status and future directions.
[13] Ling Zhou, Stephen Yau, "Efficient association rule mining among both frequent and infrequent items" Computers and Mathematics with Applications 54 (2007) 737–749.
[14] Mehdi Adda1, Lei Wu2, Sharon White2, Yi Feng 3" Pattern detection with rare item-set mining" International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.
[15] Alex tze hiang sim, Maria indrawan, Bala Srinivasan" Mining Infrequent and Interesting Rules from Transaction Records" 7th WSEAS Int. Conf. on artificial intelligence, knowledge engineering and data bases (AIKED'08), University of Cambridge, UK, Feb 20-22, 2008.
[16] Laszlo Szathmary1, Petko Valtchev1, and Amedeo Napoli2" Generating Rare Association Rules Using the Minimal Rare Itemsets Family" International Journal of Software and Informatics, ISSN 1673-7288 Vol.4, No.3, September 2010, pp. 219–238.
[17] Luigi Troiano, Giacomo Scibelli, Cosimo Birtolo "A Fast Algorithm for Mining Rare Itemsets" 2009 Ninth International Conference on Intelligent Systems Design and Applications.
[18] Laszlo Szathmary1, Petko Valtchev2, Amedeo Napoli3, and Robert Godin2 2012." Efficient Vertical Mining of Minimal Rare Itemsets" ISBN 978{84{695{5252{0, pp. 269{280, 2012.

[19] Ashish Gupta Akshay Mittal Arnab Bhattacharya 2011 "Minimally Infrequent Itemset Mining using Pattern-Growth Paradigm and Residual Trees" 17th International Conference on Management of Data COMAD 2011, Bangalore, India, December 19–21, 2011 Computer Society of India, 2011.
[20] Saha, Budhaditya, Lazarescu, Mihai and Venkatesh, Svetha 2007, Infrequent item mining in multiple data streams, in Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on; ICDM 2007, IEEE, Omaha, NE, pp. 569-574.
[21] Xindong wu, Shichao zhang" Efficient Mining of Both Positive and Negative Association Rules" ACM Transactions on Information Systems, Vol. 22, No. 3, July 2004, Pages 381–405.
[22] Johanbjarnle , Peterzhu "Mining Infrequent Patterns" Linköping University, 2009 Tnm 033 data mining.
[23] weka, http://www.cs.waikato.ac.nz/ml/weka/, 2009.
[24] S. Muthukrishnan, R. Shah, J. Vitter, "*Mining Deviants in Time Series Data Streams"*, Proceedings of the 16th International Conference on Scientific and Statistical Database Management, Santorini Island, Greece, pp.41-50, 2004.
[25] Ahmedur Rahman, C.I. Ezeife, A.K. Aggarwal "WiFi Miner: An Online Apriori-Infrequent Based Wireless Intrusion Detection System"2012.

AUTHORS

K. Sujatha is pursuing her Ph.D. in Krishna University, Machilipatnam, A.P. She is interested doing research in data mining. She has two international journal publications in data mining one with impact factor. She has a total of 9 years experience in teaching. She is working as Associate Professor in CSE Department, Malineni Lakshmaiah Engineering College, Singaraya konda, Prakasam District. A.P.

K. Rajasekhara Rao is a Professor of Computer Science & Engineering at K. L. University and presently holding several key positions in K. L. University, as Dean (Administration) & Principal, K L College of Engineering (Autonomous). Having more than 26 years of teaching and research experience, Prof. Rao is actively engaged in the research related to Embedded Systems, Software Engineering and Knowledge Management. He had obtained Ph.D in Computer Science & Engineering from Acharya Nagarjuna University (ANU), Guntur, Andhra Pradesh and produced 58 publications in various International/National Journals and Conferences. Prof.KRR was awarded with "Patron Award" for his outstanding contribution, by India's prestigious professional society Computer Society of India (CSI) for the year 2011 in Ahmedabad. He has been adjudged as best teacher and has been honored with "Best Teacher Award", seven times. Rajasekhar is a Fellow of IETE, Life Member's of IE, ISTE, ISCA & CSI (Computer Society of India). Dr. Rajasekhar is nominated as sectional committee member for Engineering Sciences of 100th Annual Convention of Indian Science Congress Association. He has been the past Chairman of the Koneru Chapter of CSI.